Erkenn
DOI 10.1007/s10670-016-9868-9

ORIGINAL RESEARCH

The Explication Defence of Arguments from Reference

Mark Pinder[1]

Received: 11 January 2016 / Accepted: 26 December 2016

Abstract In a number of influential papers, Machery, Mallon, Nichols and Stich have presented a powerful critique of so-called arguments from reference, arguments that assume that a particular theory of reference is correct in order to establish a substantive conclusion. The critique is that, due to cross-cultural variation in semantic intuitions supposedly undermining the standard methodology for theorising about reference, the assumption that a theory of reference is correct is unjustified. I argue that the many extant responses to Machery et al.'s critique do little for the proponent of an argument from reference, as they do not show how to justify the problematic assumption. I then argue that it can in principle be justified by an appeal to Carnapian explication. I show how to apply the explication defence to arguments from reference given by Andreasen (for the biological reality of race) and by Churchland (against the existence of beliefs and desires).

1 Introduction

In a number of influential papers, Machery, Mallon, Nichols and Stich have developed a powerful critique of what they call *arguments from reference*—arguments that assume a theory of reference in order to establish a substantive conclusion (Machery et al. 2004, 2013; Mallon et al. 2009). This style of argument is often used in debates about whether, for some *F*, there are *F*s. For example, one might argue that there are no beliefs, races, moral properties, colours, etc., because (say) nothing satisfies the descriptions that common-sense associates with the

✉ Mark Pinder
mark.jonathan.pinder@gmail.com

[1] University of Birmingham, Birmingham, UK

Published online: 10 February 2017

corresponding terms—thereby assuming a descriptivist theory of reference—;[1] or one might argue that there *are* beliefs, races, moral properties, colours, etc., because (say) there is a common essence amongst the things we typically use those terms to denote—thereby assuming a causal-historical theory of reference.[2] Machery et al. critique such arguments by attacking the assumption that a particular theory of reference is correct: they argue that the standard methodology for theorising about reference is flawed, and thus that the assumption of any theory of reference in an argument from reference is unjustified. They conclude that every argument from reference has an unjustified assumption.

The aim of the present paper is twofold. First, I argue that the many extant responses to Machery et al.'s critique do little for the proponent of an argument from reference, as the responses do not show how to justify her assumption of a theory of reference. If philosophers are to continue using arguments from reference, an acceptable strategy for justifying the assumption of a theory of reference must be provided.

Second, I show how the assumption of a theory of reference in an argument from reference can in principle be justified by an appeal to *explication*.[3] In particular, the assumption of a theory of reference for the relevant common-sense terms can be justified by an explication of the common-sense concepts expressed by those terms. I call this response the *explication defence* of arguments from reference. By way of illustration, I apply the explication defence to two of Machery et al.'s principal examples of arguments from reference: Andreasen's (2000) argument for the biological reality of race; and Churchland's (1981) argument against the existence of beliefs, desires and other familiar mental states. (Towards the end of the paper, I also briefly suggest that the defence can be applied to Boyd's (1988) defence of moral knowledge as a posteriori.) In appropriate contexts, then, arguments from reference may be sound.

2 Background

A *theory of reference* seeks to provide a systematic basis upon which worldly items are assigned to some collection of words or concepts. Broadly speaking, there are two principal types of theory of reference: descriptivist theories and causal-historical theories. *Descriptivist theories of reference* hold that competent speakers associate *reference-fixing descriptions* with the relevant terms or concepts, and that the referent of a given term or concept is whatever satisfies, or comes sufficiently close to satisfying, the relevant reference-fixing descriptions.[4] Thus, on a

[1] E.g., respectively: Churchland (1981), Appiah (1985), Mackie (1977) (viz. the argument from queerness), Maund (2011).

[2] E.g., respectively: Lycan (1988), Andreasen (2000), Boyd (1988), and (perhaps) Bradley and Tye (2001).

[3] Schupbach (2015) and Shepherd and Justus (2015) argue that experimentation can play a positive role in explication. The general idea in the present paper is quite different, and is not in tension with these authors' work.

[4] I stay neutral throughout on the distinction between *strong* descriptivism, on which the reference-fixing description also constitutes the *meaning* of the term, and *weak* descriptivism, on which the reference-fixing description merely provide a *referent* for the term.

descriptivist theory for "Gödel", if Padma associates the description *proved the incompleteness theorem* with "Gödel", and if Gödel proved the incompleteness theorem, then (in Padma's language) "Gödel" refers to Gödel. In contrast, *causal-historical theories of reference* hold that a term or concept t refers to an entity x just in case there is an appropriate causal-historical chain of users acquiring t from other users, such that the chain began with an initial 'baptism' of x. Thus, on a causal-historical theory for "Gödel", if Padma's use of "Gödel" can be traced back via an appropriate causal-historical chain (of other users of "Gödel") to a 'baptism' of Gödel, then (in Padma's language) "Gödel" refers to Gödel.

Machery et al. characterise *arguments from reference* as "arguments that derive philosophically significant conclusions from the assumption of one or another theory of reference" (Mallon et al. 2009: 332). Herein, I principally focus on two of their principal examples.[5]

The first is an argument drawn from Andreasen 2000 in favour of the biological reality of race. Andreasen begins by offering a 'cladistic definition of race' in terms of a phylogenetic classification of human breeding populations, where "[a] 'breeding population' is a set of local populations that exchange genetic material through reproduction and are reasonably reproductively isolated from other such sets" (p. S659). The idea is that, when various subpopulations of a 'parent' breeding population separate from each other—perhaps for geographical or socio-cultural reasons—'daughter' breeding populations are generated. Andreasen then cites research (Cavalli-Sforza 1991) that plausibly supports the conclusion that such cladistic populations are in fact biologically real. Starting from the "undifferentiated stock of modern humans evolving in Africa $\sim$200,000 years ago" (2000: S660),

> [t]he first split divides Africans from all other populations. The second split represents a division between Pacific-Southeast Asians and the rest of the world. After that division, the Australopapuans diverged from the rest of Pacific-Southeast Asia, and the fourth split separates northeast Asians and Amerindians from European and non-European Caucasoids. (*ibid.*)

Andreasen takes this to support the conclusion that race is biologically real.

Andreasen is explicit that her cladistic definition of race deviates from common-sense conceptions of race (pp. S661–S662); she tentatively suggests that, according to common-sense, "races are demarcated by appeal to observable properties (e.g., skin color, hair type, and eye shape) [that may be taken to be] good predictors of more significant inherited differences (e.g., behavioural, intellectual, or physiological differences)" (p. S663). To justify this deviation, Andreasen points to causal-historical theories of reference:

> The objectivity of a kind, biological or otherwise, is not called into question by the fact that ordinary people have mistaken beliefs about the nature of that kind. Those familiar with the causal theory of reference for natural kind terms will be aware of this possibility. (p. S662)

[5] See Mallon et al. (2009: 333–337).

According to Machery et al., Andreasen here assumes a causal-historical theory of reference for "race".[6]

An anonymous reviewer makes the helpful point that there are essentially two components to Andreasen's argument: a metaphysical/scientific sub-argument that the relevant cladistic populations are biologically real; and a semantic sub-argument drawing upon theories of reference to establish that "race" can be used to denote such populations. Importantly, *both* components are required to establish Andreasen's conclusion that race is biologically real.[7] Andreasen does not merely seek to establish that cladistic populations are biologically real, subsequently choosing to express this conclusion using "race": the conclusion is rather that those biologically real populations *are* races—that races, not merely cladistic populations, are a part of biological reality. In this sense, Andreasen's conclusion is a substantive conclusion concerning human biology and race.

Thus, according to Machery et al., Andreasen assumes a causal-historical theory of reference in order to establish the substantive conclusion that race is biologically real.

The second example is an argument by Churchland (1981) in support of his eliminative materialism. Churchland begins by arguing that the common-sense mental-state concepts of belief, desire, etc., are theoretical concepts of folk psychology. Here, folk psychology is construed as a theory consisting of a set of laws. Amongst these laws are those that either: (1) connect external circumstances to mental states; (2) connect mental states to other mental states; or (3) connect mental states to actions. For example, respectively:

(1) For all x, y: if y is in x's field of vision, then (ceteris paribus) x is aware of y.
(2) (i) For all x, p: if x fears p, then (ceteris paribus) x desires not-p.
(ii) For all x, p, q: if x believes p and x believes [p implies q], then (ceteris paribus) x believes q.
(3) For all x, p, φ: if both x desires p, and x believes that φing will bring about that p, then (ceteris paribus) x will φ.[8]

The theoretical concepts of folk psychology (with which Churchland identifies the common-sense mental-state concepts) are just the concepts expressed by the non-logical vocabulary in such laws. Churchland goes on to argue that contemporary science shows that it is a serious possibility that folk psychology is radically false. He thus concludes that it is a serious possibility that there are no beliefs, desires, etc.

This argument assumes that the theoretical concepts of folk psychology have their extensions fixed descriptively, with the laws of folk psychology acting as reference-fixing descriptions. It is only with this assumption that the radical falsity

[6] See Mallon et al. (2009: 337). *Pace* Mallon et al., Andreasen does not commit herself to the assumption of a causal-historical theory of reference. Rather, she claims that her "point does not depend on the specifics of the causal theory", and gives a few examples "to support the idea that the objectivity of a kind is not undermined by the fact that ordinary people have mistaken beliefs about its nature" (2000: S662). The discussion in Sect. 4 provides one way of filling in the details of Andreasen's line of thought.

[7] Note that this conclusion is essential to Andreasen's aims: she ultimately seeks to establish that biological realism and social constructivism are *compatible* views about *race* (2000: S653–S655).

[8] For more examples, see Churchland (1979: ch. 4).

of folk psychology would suffice to establish that there are no beliefs, desires, etc. In contrast, given (say) a causal-historical theory of reference for those concepts, Churchland's argument is invalid: the theoretical concepts of folk psychology might stand at the end of appropriate causal-historical chains, at the other end of which are successful baptisms, *even if* the laws of folk psychology are radically false of whatever was baptised. In that case, even if common-sense mental-state concepts are theoretical concepts of folk psychology, the radical falsity of the laws of folk psychology would not suffice to establish that there are no beliefs, desires, etc.

It is worth emphasising that, like Andreasen, Churchland seeks a *substantive* conclusion. In particular, Churchland seeks to draw a substantive conclusion about human psychology and mental architecture: whatever psychological states there are in humans, belief, desire, etc., are not among them—beliefs, desires, etc., are not a part of human psychological reality. Such claims are not in any sense metalinguistic, nor are they about our intuitions concerning human psychology or mental architecture. Churchland's arguments assumes a theory of reference in support of a substantive claim concerning human psychology and mental architecture.

Machery et al.'s critique of arguments from reference is designed to undermine the assumption of a theory of reference. We can understand the critique as consisting of four claims.

Claim One: experimental evidence suggests that there is cross-cultural variation in semantic intuitions—and, in particular, intuitions about reference. Machery et al. summarise the evidence as follows.

> In two separate studies […], we found that Americans were more likely than Chinese to give causal-historical responses. […] As we had predicted, Chinese participants tended to have descriptivist intuitions, while Americans tended to have [causal-historical] intuitions. (Mallon et al. 2009: 341)

Claim Two: cross-cultural variation in semantic intuitions would undermine the use of semantic intuitions in theorising about reference. The issue is that

> philosophers usually appeal only to their own intuitions about reference and those of a few colleagues, perhaps because they take these intuitions to be representative of competent speakers' intuitions or perhaps because they take them to be more reliable. (Machery et al. 2013: 620)

However, according to Machery et al.,

> the evidence suggests that it is wrong for philosophers to assume a priori the universality of their own semantic intuitions. […] We find it wildly implausible that the semantic intuitions of the narrow cross-section of humanity who are Western academic philosophers are a more reliable indicator of the correct theory of reference (if there is such a thing […]) than the differing semantic intuitions of other cultural or linguistic groups (Machery et al. 2004: B8–B9)

Claim Three: semantic intuitions would be required to determine which theory of reference (*if any*) *is correct*. The idea is that,

> [o]nce we see that lots of assumptions are needed to turn [a] "picture" into a full-fledged theory [of reference], we realize, first, that there are lots of nearby theories of reference, and, second, that jiggling one or another fine point […] yields an alternative theory from which different philosophically significant conclusions follow. […] It seems that in practice the way to know which one of these options is the right one is to rely on intuitions [about reference] […]. (Machery et al. 2013: 624)

Claim Four: the assumption of a theory of reference (*in any argument from reference*) *is unjustified if we are not in a position to determine whether or not it is correct*. So without an acceptable methodology for theorising about reference, the assumption of any particular theory of reference is unjustified.

Machery et al. conclude that, given the available experimental evidence, every argument from reference has an unjustified assumption: as they stand, all arguments from reference fail.

Let me make two immediate comments about the critique. First, the particular intuitions tested by Machery et al. concern proper names, whereas the arguments from reference they cite make use of terms—such as "race" and "belief"—that are *not* proper names.[9] However, Machery et al. take it that cross-cultural variation in folk semantic intuitions about proper names may indicate cross-cultural variation in folk semantic intuitions more generally. In what follows, I accept this for the sake of argument. (My discussion of the literature in Sect. 3 would apply *mutatis mutandis* had the literature not focused on proper names.)

Second, Machery et al.'s critique draws upon alleged variation in folk semantic intuitions to undermine arguments from reference. This may seem strange, as variation in such intuitions appears to be independent of such substantive biological and psychological issues as whether race is biologically real and whether there are beliefs and desires.[10] Nonetheless, this apparent independence does not serve to undermine the critique. Machery et al. are certainly right to think that the assumption of a theory of reference in an argument from reference requires *some kind* of justification. And this is so whether or not the conclusions of such arguments are independent of the alleged cross-cultural variation in semantic intuitions. To respond to the critique, the proponent of an argument from reference must provide a justification for her assumption of a theory of reference—and if that justification makes use of semantic intuitions, then she must explain, in light of the experimental data, why the justification stands.

[9] Perhaps "race" and "belief" are more like *kind* terms. Experimental work on kind terms, though, has not addressed the question of cross-cultural variation. See e.g., Braisby et al. (1996), Jylkkä et al. (2009), Genone and Lombrozo (2012), Nichols et al. (2016).

[10] To be clear, Machery et al. would probably agree with this point. E.g., Mallon (2006: 547ff) argues that disputes about the correct theory of reference for "race" do not help us resolve important metaphysical questions in the philosophy of race.

3 Extant Responses

The critique has given rise to a great deal of discussion. Typically, this discussion has not explicitly focused on whether Machery et al. have successfully undermined arguments from reference, but has rather focused on Claims One, Two and Three of the critique.[11,12] In this section, I consider whether this discussion provides the proponent of an argument from reference with an adequate response to Machery et al.'s critique. I argue not.

3.1 Responses to Claim One

A number of theorists argue that Machery et al.'s experimental data fail to show that there is cross-cultural variation in folk semantic intuitions.

In Machery et al. (2004), English-speaking participants from Hong Kong and the US were presented with vignettes based on so-called 'Gödel' and 'Jonah' cases made famous by Kripke (1980). For example, in the vignette based on the Gödel-case: John has learnt that Gödel proved the incompleteness theorems, but has heard nothing else of Gödel; and, unbeknownst to John, Schmidt in fact did the work in question but died in mysterious circumstances, after which Gödel got hold of the manuscript and took credit for the work. After each vignette, participants were asked a question, designed to elicit either a descriptivist intuition or a causal-historical intuition; for the Gödel-case:

When John uses the name "Gödel", is he talking about:

(A) the person who really discovered the incompleteness of arithmetic? or
(B) the person who got hold of the manuscript and claimed credit for the work?

Machery et al. found that, in response to Gödel-cases, Chinese participants were more likely to answer (A) whereas Western participants were more likely to answer (B). The authors gloss these results by stating that "Chinese participants tended to have descriptivist intuitions, while Westerners tended to have Kripkean [causal-historical] ones" (p. B7). And they take this result to support the claim that Westerners are "more likely than the Chinese to have intuitions that fall in line with causal-historical accounts of reference" (p. B8)—i.e. that there is cross-cultural variation in folk semantic intuitions.

A number of objections have been raised.[13] For example, Martí (2009) claims that Machery et al.'s experiment tests *metalinguistic* intuitions about theories of reference, rather than *semantic* intuitions; and thus that the experiment fails to establish that there is cross-cultural variation in folk semantic intuitions. In contrast, Deutsch (2009: 453ff) claims that the experiment fails to distinguish between intuitions about the *speaker* reference of "Gödel" (i.e. the object to which John

[11] One exception is Ichikawa et al. (2012: 67–68). Their brief comments rely on their response to claim three of Machery et al.'s critique, which I discuss in Sect. 3.3 below.

[12] Cf. Hansen (2015).

[13] In addition to those discussed below, Lam (2010) criticises the lack of Cantonese vignettes; see Machery et al. (2010) for a response.

intended to refer) and intuitions about the *semantic* reference of "Gödel" (i.e. the object assigned to it by linguistic convention); and thus that the experiment fails to establish that there is cross-cultural variation in folk intuitions concerning *semantic* reference. Sytsma and Livengood (2011) claim that Machery et al.'s vignettes contain an ambiguity in epistemic perspective; the experiments fail to distinguish between whether the participants are answering the question from *John*'s epistemic perspective or from *the narrator*'s epistemic perspective. Systma and Livengood suggest that participants might conceivably answer (A) if they are limiting themselves to the information available to John, even if those participants would answer (B) given the information available to the narrator. Sytsma and Livengood conclude that, as Machery et al.'s data could be explained by cross-cultural variation in how epistemic perspective is disambiguated, the experiment fails to establish that there is cross-cultural variation in folk semantic intuitions.

In light of such objections to Machery et al.'s experimental work, the proponent of an argument from reference might reason as follows: *the experimental evidence does not support the conclusion that there is cross-cultural variation in folk semantic intuitions, so Claim One is false, so Machery et al. have failed to show that the assumption of a theory of reference in arguments from reference is unjustified*.

There are at least three problems with this line of reasoning, however. First, further experimental work suggests that Machery et al.'s findings are in fact reasonably robust. Revised versions of the initial experiment designed to take into account the objections of Martí, Deutsch and Sytsma and Livengood appear to successfully replicate the initial results. See, respectively, Machery et al. (2009, 2015), and Sytsma et al. (2015). The increasing stock of data is by-and-large consistent with the hypothesis that there is cross-cultural variation in semantic intuitions.

Second, even if it were shown that there is no cross-cultural variation in folk semantic intuitions, this would not suffice as a response to Machery et al.'s critique. As Machery et al. make clear (e.g., 2013: 632–633), it is not enough that one's assumption of a theory of reference has not been shown to be unjustified; the proponent of an argument from reference still owes us an account of how to collate the relevant intuitions to yield "a full-fledged theory of reference (causal-historical or otherwise) that can serve to underwrite arguments from reference" (Machery et al. 2013: 633). And this is not straightforward: "resourceful descriptivists can accommodate the intuitions about actual cases that form the basis of [Kripke's arguments] (just as resourceful non-descriptivists can accommodate the intuitions about 'Madagascar' and 'King Arthur' […])" (p. 625). It would remain incumbent upon the proponent of an argument from reference to provide a justification for the particular theory of reference that underpins her argument from reference.

Third, regardless, if the viability of an appeal to semantic intuitions depends on whether or not there is cross-cultural variation in folk semantic intuitions, then those intuitions are ill-suited for justifying the assumption of a theory of reference in arguments from reference. Recall that arguments from reference seek to establish substantive conclusions about, for example, whether race is biologically real and whether there are beliefs and desires. Yet, as suggested above, variation in folk semantic intuitions is not relevant to such substantive biological and psychological

issues: there is no reason to think that the biological reality of race, or facts about our mental architecture, are so straightforwardly sensitive to the distribution of folk semantic intuitions. If Andreasen's and Churchland's arguments from reference are at all plausible, then the arguments should not depend on whether there is cross-cultural variation in semantic intuitions. That is, the proponent of an argument from reference would like a justification for her assumption of a theory of reference such that the justification is independent of the distribution of folk semantic intuitions. Preferably, the justification for the assumption of a theory of reference in an argument from reference should rely on factors that are plausibly relevant to the substantive conclusion of the argument from reference. The proponent of an argument from reference, then, should not rely on the possibility that there is no cross-cultural variation in folk semantic intuitions.

3.2 Responses to Claim Two

Some theorists argue that, even if there is cross-cultural variation in folk semantic intuitions, this does not suffice to undermine the use of semantic intuitions for theorising about reference. According to the *expertise defence*, the intuitions of experts-about-reference carry more weight than folk intuitions, and are not undermined by Machery et al.'s experimental data. I focus here upon Devitt's variant of the defence.[14]

According to Devitt, semantic intuitions are *theory-laden* in an important sense.

> [T]he intuitions are mostly *the product of experiences* of the linguistic world. They are like "observation" judgments. As such, they are "theory-laden" in just the way that we commonly think observation judgments are. (Devitt 2012a: 19)

In particular, for Devitt,

> Linguistic education should make a person a better indicator of linguistic reality just as biological education makes a person a better indicator of biological reality. (2006: 115)

Thus, for example, intuitions about reference in a given case—such as a Gödel-case—may be affected by one's belief in a given theory of reference. Devitt concludes that

> we should prefer the linguistic intuitions of linguists and philosophers because they have the better background theory and training. (2012a: 19)

If this is right, then the philosophical import of Machery et al.'s experimental data is greatly diminished; the variation in folk semantic intuitions in Gödel-cases may simply be insignificant next to the intuition, widespread amongst experts-about-reference, that "Gödel" refers to the publisher of the manuscript.

[14] See Devitt (2011, 2012a, b), and Machery et al. (2013), Machery (2012a, b) for responses. For other variants, see e.g., Cohnitz and Haukioja (2015), Ludwig (2007).

Inspired by the expertise defence, the proponent of an argument from reference may reason as follows: *the intuitions of experts-about-reference carry more weight than folk intuitions, so Machery et al.'s empirical work fails to undermine the use of semantic intuitions in theorising about reference, so Claim Two can be rejected, so Machery et al. have failed to show that the assumption of a theory of reference in arguments from reference is unjustified.*

In parallel to the discussion in the previous subsection, there are three problems with this line of reasoning. First, Machery presents experimental data that threatens the expertise defence (2012a: 45–52). Roughly, the data suggest that, if expertise about reference has an effect upon intuitions, then the effect is *inconsistent* across different disciplines: semanticists and philosophers of language tended to have more causal-historical intuitions than similarly educated lay people, whereas discourse analysts, sociolinguists and historical linguists tended to have more descriptivist intuitions than similarly educated lay people.[15] While this study is tentative and more work is required—for example, in Machery's study, the latter tendency did not reach statistical significance—we should be cautious about the claim that expertise about reference increases the reliability of intuitions about reference.

Second, even if the semantic intuitions of experts-about-reference *do* carry more weight than folk semantic intuitions, this does not *per se* provide a justification for the assumption of a theory of reference in arguments from reference. As before, the proponent of an argument from reference would have to show how those intuitions support whichever theory of reference underwrites her argument from reference.

Third, it is not clear that a plausible argument from reference should depend in this way on the semantic intuitions of experts-about-reference. Arguments from reference seek to establish substantive conclusions about, for example, biology and psychology. And the questions of whether race is biologically real, or of whether there are beliefs, desires, etc., are not obviously sensitive to the semantic intuitions of experts-about-reference. With respect to substantive biological and psychological questions, it is not particularly plausible that the semantic intuitions of Devitt and Kripke should carry so much more weight than, say, those of biologists and psychologists. In general, the proponent of an argument from reference would like to provide a justification for her assumption of a theory of reference *such that the justification does not rely pivotally on the semantic intuitions of experts-about-reference*. Preferably, the justification for the assumption of a theory of reference in an argument from reference should rely on factors that are plausibly relevant to the substantive conclusion of the argument from reference. The proponent of an argument from reference, then, should not rely on the expertise defence.

[15] Should the latter group of theorists count as experts about reference? Machery (2012a: 51–52) claims so, but provides minimal argument; Devitt (2012a: 24) doubts Machery's conclusion without (as far as I can tell) addressing Machery's argument. In at least this respect, more work is required before we can confidently interpret Machery's findings. Regardless, we cannot simply *assume* that discourse analysts, sociolinguists and historical linguists are *in*expert about reference.

3.3 Responses to Claim Three

A number of theorists have argued that Machery et al. overestimate the importance of semantic intuitions and, in particular, intuitions about Gödel-cases.[16] For example, Ichikawa et al. (2012) and Devitt (2011) downplay the importance of *semantic* intuitions. Ichikawa et al. emphasise

> Kripke's first argument […] that for most names, most users of the name cannot give an individuating description of the bearer of the name. […] The best they can do for 'Cicero' is 'a famous Roman orator' and the best they can do for 'Feynman' is 'a famous physicist'. (2012: 61)

See Kripke (1980: 81–82). As Ichikawa et al. understand Kripke, this argument establishes that *some* names do not have their referents fixed descriptively; the Gödel-case is merely intended "to show that the number of descriptive names in English is not just small, it is *very* small" (p. 63). Devitt (2011: 421–424), on the other hand, emphasises the importance of *modal* intuitions—such as that Feynman could have not been a physicist—which appear to be incompatible with (strong) descriptivist theories of reference.

In contrast, Deutsch (2009: 450f) seeks to downplay the importance of intuitions in general; he claims that it is not *the intuition that* "Gödel" *refers to Gödel* that is intended to undermine the descriptivist theory of reference, but *that* "Gödel" *refers to Gödel*. And Deutsch suggests (pp. 451–452) that Kripke has independent arguments for that the claim that "Gödel" refers to Gödel. Here is Deutsch's first example.

> All that many of us 'know' about Peano is that he was the discoverer of certain axioms concerning the natural numbers. But it turns out that Dedekind discovered those axioms. If descriptivism is true, many of us have been referring all along to Dedekind with our uses of 'Peano'. But we have not been referring to Dedekind with those uses. We have been referring instead to Peano, *mis*attributing to him the discovery of the axioms. This […] strengthens the claim that the Gödel-case is a counterexample [to descriptivism] by showing us that the way in which we *ought* to judge, with respect to the imaginary Gödel-case, should line up with the way in which we do in fact, and correctly, judge about the real-life Peano case. (pp. 451–452)

Such arguments, according to Deutsch, do not appeal to intuitions.

Inspired by such responses, the proponent of an argument from reference may reason as follows: *semantic intuitions are not necessary for determining which theory of reference* (*if any*) *is correct, so Claim Three is false, so Machery et al. have failed to show that the assumption of a theory of reference in arguments from reference is unjustified.*

Two of the problems discussed above apply to this line of reasoning. First, it does not suffice as a response to Machery et al. to downplay the importance of semantic intuitions in theorising about reference. As noted above, it would remain incumbent

[16] In addition to the discussion below, see Devitt (2011: 420–424).

upon the proponent of an argument from reference to provide a justification for her particular assumption of a theory of reference.

Second, such substantive issues as the biological reality of race and human mental architecture are independent of the data discussed by Ichikawa et al., Devitt and Deutsch; they are independent of whether most users of "Feynman" have an individuating description of Feynman, whether Feynman could have not been a physicist, and whether Peano axiomatised the natural numbers. If Andreasen's and Churchland's arguments from reference are plausible at all, the justification for their assumptions of theories of reference should be independent of the kinds of data discussed by Ichikawa et al., Devitt and Deutsch. Ultimately, I suggest, the proponent of an argument from reference should seek an altogether different kind of justification for her assumption of a theory of reference.

4 The Explication Defence of Arguments from Reference

The assumption of a theory of reference in an argument from reference can, in principle, be justified by explication. Although this *explication defence* may not work in all cases, it is effective for at least two of Machery et al.'s principal targets: Andreasen's and Churchland's arguments from reference.

4.1 Explication

Explication is to be understood, in a broadly Carnapian sense, as a stipulated refinement of a common-sense or otherwise imprecise concept, in order to facilitate subsequent theorising.[17] The common-sense or imprecise concept with which we begin is the *explicandum*, and the refined concept is the *explicatum*.

Explications are *stipulative*: an explicatum is neither intended to *describe* the explicandum nor encode its ordinary usage. Rather, the explicatum is offered as a theoretical replacement of the explicandum. When theorising, so goes the thought, one *ought* to use the explicatum *in place of* the explicandum.

There may be good explications and bad explications. Whether an explication is good will depend on various criteria. Carnap provides four criteria, which an explication should satisfy to "a sufficient degree" (1950: 7).

(I) The explicatum should be similar in relevant respects to the explicandum.
(II) The explicatum should be precise.
(III) The explicatum should be a fruitful concept.
(IV) The explicatum should be simple.

With respect to (I), Carnap notes that "close similarity is not required, and considerable differences are permitted" (*ibid.*). The extension of the explicandum need not be preserved, so long as *some* key features are preserved. With respect to (II), explicit rules for using the explicatum should be given; these rules can, but need not, be given in the form of a definition. With respect to (III), Carnap

[17] See Carnap (1950).

understands a concept to be *fruitful* insofar as it is "useful for the formulation of many universal statements" (*ibid*.), where these statements should presumably underpin relevant explanations and predictions. And, finally, (IV) is taken to be subordinate to (I)–(III); as such, I largely leave it to one side in what follows.

Let me provide an immediate example. Consider the concept of *planet*.[18] Until recently, there was no agreed upon definition—merely nine canonical instances. Then, in 2006, the International Astronomical Union explicated the concept, refining it so as to provide a better taxonomy of celestial objects. A *planet* was henceforth to be an object such that: (a) it orbited a star but did not orbit another planet; (b) it was large enough for gravity to have formed it into a sphere but not large enough for its gravity to trigger fusion; and (c) it had cleared its orbit of debris. Here, the explicandum is the pre-2006 concept, with its nine canonical instances; and the explicatum is the refined concept with its tripartite satisfaction condition.

The explication of the concept of planet plausibly satisfies Carnap's criteria. Criterion (I) is satisfied: eight of the nine canonical instances of the explicandum—all except Pluto, which fails to satisfy (c)—fall under the explicatum; and objects that would previously have been deemed canonical non-planets—such as the moon, the sun, shards of ice in the Kuiper Belt, etc.—do not fall under the explicatum. Criterion (II) is satisfied: the explicatum is introduced with a tripartite definition (a)–(c). Criterion (III) will plausibly be shown to be satisfied in due course: insofar as the explicatum provides a unified cosmological kind, it is likely that the explicatum will prove to be fruitful—or, at least, *more* fruitful than the explicandum. And, insofar as the tripartite definition (a)–(c) is simple, the explication satisfies criterion (IV). By Carnap's lights, then, the IAU's explication of the concept of planet was plausibly a good explication.

There are two features of this account of explication that will play an important role in what follows. The first feature:

> *Referential Control*. The theorist can stipulate how the extension of an explicatum is fixed. That is, the theorist can stipulate which *theory of reference* applies to an explicatum.

Referential Control follows jointly from the fact that explications are stipulative, and the fact that the extensions of explicanda need not be preserved. By way of example, consider again the IAU's explication of *planet*. It is natural to understand the IAU as implicitly stipulating that a descriptivist theory of reference applies to the explicatum: its extension contains whatever comes sufficiently close to satisfying (a)–(c).

The second feature is this:

> *Explicandum Replacement*. When theorising about a subject matter, one ought to use appropriate explicata *in place of* the corresponding explicanda.

[18] I roughly follow Ludlow's (2014: 41ff) discussion here, although Ludlow is concerned with meaning modulation rather than explication.

Here, I take *appropriate* explicata to be those that (i) are relevant to the subject matter and (ii) are the result of *good* explications.[19] How do we use an explicatum *in place of* an explicandum? Suppose that an ordinary term t ordinarily expresses a common-sense concept c, and that a good explication of c yields the explicatum c_E. Then, we use c_E in place of c by taking t to express c_E.

By way of example, consider again the IAU's explication of *planet*. Suppose that, as I suggested above, the explication was a *good* explication. Then, according to Explicandum Replacement, a cosmologist (qua cosmologist) should now take "planet" to express the explicatum in place of the explicandum. That is, when theorising, the cosmologist should *ceteris paribus* be willing to assert "Pluto is not a planet", and to deny "Pluto is a planet".

Before proceeding, I note that a variety of philosophical objections have been raised against Carnap's account of explication. Carnap and others have responded to these objections.[20] I will not re-address such objections, but will assume that explication is a legitimate method for concept refinement. On that assumption, I argue that the proponent of an argument from reference can justify her assumption of a theory of reference. For those who do not endorse explication, the following discussion nonetheless provides an example of the *kind* of justification that the proponent of an argument from reference must provide for her assumption of a theory of reference: given something like Referential Control and Explicandum Replacement, she can adequately respond to Machery et al.'s critique.

4.2 Andreasen's Argument

Explication can be used to justify the assumption of a theory of reference in arguments from reference. We begin with Andreasen's argument, introduced in Sect. 2, which points to causal-historical theories of reference in justifying deviation from common-sense conceptions of race. According to Machery et al., the argument thus assumes a causal-historical theory of reference for "race" in order to establish the substantive conclusion that race is real.

Now, *pace* Machery et al., I do not think that Andreasen's argument assumes a *causal-historical* theory of reference. Rather, as I construe the argument, it in fact assumes a *descriptivist* theory of reference. So let me explain how we can understand the argument in terms of an explication and, afterwards, I will explain why Machery et al. characterise Andreasen to be assuming a causal-historical theory of reference.

Andreasen's argument is naturally construed as built upon an explication. The explicandum is the common-sense concept of race. The explicatum is introduced as

[19] This issue is complicated if we acknowledge that, for a given explicandum, there may be competing explicata. In particular, *multiple* explicata may be (i) relevant to the subject matter and (ii) the result of good explications. There is latitude about what we say about such a situation; perhaps the theorist is permitted to use either explicata or perhaps she is required to use the *best* explicatum. (The former option would complicate the explication defence, but is compatible with the general strategy.) For simplicity, I leave this issue to one side throughout.

[20] See e.g., Carnap (1963), Justus (2012), Kitcher (2008), Maher (2007).

follows. (The details of the proposal are not important for present purposes; see Andreasen 1998, 2000 for a more comprehensive account.)

> Although the principles of cladistics classification were developed for defining higher taxa, they can be adapted for defining race. A cladistic view of race would require constructing a phylogenetic tree out of human breeding populations; the nodes would represent breeding populations and the branches would represent the births of new breeding populations. A 'breeding population' is a set of local populations that exchange genetic material through reproduction and are reasonably reproductively isolated from other such sets. […] A breeding population is 'born' when a local subpopulation becomes separated for its parent population and there is limited gene flow between "parent" and "offspring". […] The terminal nodes represent current breeding populations, the whole tree represents the human species, and the nested hierarchy of monophyletic units represents a nested hierarchy of races. (2000: S659)

More succinctly, the explicatum is defined as a kind of subpopulation of humans, such that the subpopulation is represented by a monophyletic unit in an appropriate cladistic classification of human breeding populations.

Now, by Referential Control, Andreasen can stipulate a theory of reference for her explicatum. And, quite clearly, she stipulates a *descriptivist* theory of reference: the extension of the explicatum contains whatever satisfies the above definition. (This is clear, for example, on pp. S659–S661.)

By Explicandum Replacement, if this explication is a good explication, then one should use the explicatum *in place of* the explicandum for the purpose of theorising about human biology. So let us turn to Carnap's criteria.

Criterion (I) demands similarity between explicandum and explicatum. Andreasen argues that

> there are at least two important elements of [the common-sense conception of race] that the cladistic concept retains. First, many people believe that races are subspecies; they are biologically objective categorical subdivisions of Homo sapiens. Second, shared ancestry has played, and probably continues to play, an important role in the ways that ordinary people think about race. (2000: S665)

If Andreasen is right, then her explication satisfies (I).

Criterion (II) demands precision. And, as Andreasen's definition is given in terms of cladistics, it is plausibly *as* precise as other cladistically-defined taxa—perhaps such as *genus* or, more controversially, *species*. So let us accept that the explicatum satisfies (II).

Criterion (III) demands fruitfulness, construed in particular as usefulness for the formulation of universal statements. And, given that the explicatum is defined in terms of reproductively isolated histories, it is likely to feature in laws about, for example, gene frequencies within and between various subpopulations, and the relatedness between individuals belonging to the same or different subpopulations. (See Andreasen 2000: S659–S660.) So, plausibly, the explicatum is fruitful.

Criterion (IV) demands simplicity. Now, it is unclear whether or not the explicatum is simple but, as criterion (IV) is subordinate, I put it aside.

So let us accept that Andreasen's explication is a good explication. Then, by Explicandum Replacement, one should use the explicatum *in place of* the explicandum for the purpose of theorising about human biology. And, as Andreasen's argument from reference is concerned (in part) with human biology, her use of "race" in that argument should express the explicatum in place of the explicandum. That is, she should take "race" to denote the kind of subpopulation that satisfies her cladistic definition.

It follows a fortiori that, if Andreasen is using "race" as she ought, then she is justified in assuming that "race" denotes the kind of subpopulation that satisfies her cladistic definition. That is, she is justified in assuming a descriptivist theory of reference for "race", where her definition provides the relevant reference-fixing descriptions. So the assumption—that the relevant theory of reference is correct—is justified.

So why do Machery et al. claim that Andreasen is assuming a causal-historical theory of reference? To answer this question, it is useful to look more abstractly at the picture I am defending. With respect to arguments from reference, both Referential Control and Explicandum Replacement play important roles. First, Referential Control allows one to stipulate a theory of reference for a *new* concept (the explicatum). Second, Explicandum Replacement permits one to use a familiar term to *express* the new concept. Together, this amounts to a justification for the assumption that a theory of reference for a familiar term is correct. As I have construed Andreasen's argument from reference, it relies on the stipulation of a *descriptivist* theory of reference for the new *concept*. However, Machery et al. focus on the assumption that "race" denotes whatever falls under Andreasen's cladistic conception of race. Given that Andreasen takes the cladistic conception of race to apply to biological kinds, combined with the fact that she mentions a Kripkean causal-historical theory of reference for natural kind terms (2000: S662), Machery et al. take this assumption to be an appeal to a causal-historical theory of reference. However, on my view, that *particular* assumption corresponds to Explicandum Replacement, which permits one to use a familiar term to express an explicatum. The explicatum itself is assumed to be subject to a *descriptivist* theory of reference—an assumption which is justified by Referential Control.

4.3 Churchland's Argument

Let us now turn to Churchland's argument: building on the claim that the common-sense mental-state concepts are theoretical concepts of folk psychology, the argument seeks to establish that, as the theoretical concepts may well be empty, it is a serious possibility that there are no beliefs, desires, etc. The argument assumes that the theoretical concepts of folk psychology are subject to a descriptivist theory of reference, where the laws of folk psychology act as reference-fixing descriptions.

Now, in light of Machery et al.'s critique, Churchland is not entitled to simply *assume* that the theoretical concepts of folk psychology are subject to a particular theory of reference. However, as we have seen, theorists may *stipulate* particular

theories of reference for their explicata. As such, it may be possible for Churchland not to *identify* common-sense mental-state concepts with theoretical concepts of folk psychology; rather, he might take himself to be *explicating* the former concepts with something like the latter.

Let me spell out the explication. The explicanda are the common-sense mental-state concepts of belief, desire, etc., that form the basis of our common-sense ideas about human psychology. The explicata are a collection of theoretical concepts, which we will take to be jointly introduced by the laws of folk psychology such as (1)–(3) above. Given Referential Control, it can be *stipulated* that these concepts are subject to a descriptivist theory of reference, with the laws of folk psychology acting as reference-fixing descriptions. Thus, their extensions contain whatever (if anything) comes sufficiently close to satisfying the laws of folk psychology. For ease of reference, I call these concepts the *descriptive theoretical concepts*.

Given Explicandum Replacement, if this explication is a good explication, then one should use the explicata *in place of* the explicanda for the purpose of theorising about human psychology. Let us see if this is a *good* explication.

Criterion (I) demands similarity between explicanda and explicata. Here is one key similarity in the present case. A distinctive feature of the explicanda—the common-sense mental-state concepts—is that they are deployed in a particular style of explanation and prediction of behaviour. (For example: she φed because she believes that p; she desires that q so she'll probably ψ; etc.) And, the explicata underpin structurally comparable explanations and predictions: the reference-fixing descriptions for the descriptive theoretical concepts are the laws of folk psychology, and those laws yield the relevant explanations and predictions of behaviour. So, in this key sense, our explicata are similar to the explicanda.

Criterion (II) demands precision. The descriptive theoretical concepts are fully specified by the laws of folk psychology. As such, full rules for their deployment are explicitly given and so, in the relevant sense, they are precise.

Criterion (III) demands fruitfulness, construed in particular as usefulness for the formulation of universal statements. In this sense, the descriptive theoretical concepts are plausibly fruitful. Recall that, for Churchland, folk psychology is construed as a set of laws. Now, on the most theoretically neutral interpretation of those laws—that is, the interpretation on which one assumes nothing more about beliefs, desires, etc., than is overtly encoded by those laws—the descriptive theoretical concepts *just are* the concepts that feature in the laws of folk psychology. So, it seems, the descriptive theoretical concepts are, in the present sense, fruitful.

There may seem to be a slight tension here. On the one hand, Churchland holds that it is a serious possibility that the laws of folk psychology are radically false; but, on the other hand, I suggest that the descriptive theoretical concepts are fruitful as they feature in the laws of folk psychology. The tension arises as it is unclear that concepts can be fruitful in virtue of featuring in laws that might well be radically false. However, concepts are often taken to be fruitful in virtue of featuring in laws that might well be radically false. For example, we might think that the technical concepts of *string* in theoretical physics, *Mendelian gene* in biology and *reference* in semantics are all fruitful concepts, despite the fact that they belong to theoretical

frameworks that might well turn out to be radically false: we have no empirical evidence in support of the existence of strings, it is unclear that there are anything like Mendelian genes, and many have argued that semantics should do without reference. Regardless, for Churchland, the resulting picture would be twofold: first, the descriptive theoretical concepts would be deemed fruitful enough for us to think that they are the appropriate explicata for our common-sense mental-state concept; second, the descriptive theoretical concepts would be deemed *in*sufficiently fruitful in light of contemporary scientific standards for us to accept that they successfully reflect psychological reality. This is a fine line, but one that is available to Churchland.

Criterion (IV) demands simplicity. Now, it is plausible that the concepts are moderately complex, as they are introduced implicitly by a network of laws rather than simple definitions. However, as criterion (IV) is subordinate to criteria (I)–(III), I again put this aside.

So let us accept that the explication in question is a *good* explication. Then, given Explicandum Replacement, one should use the explicata *in place of* the explicanda for the purpose of theorising about human psychology. That is, when theorising about human psychology, one should take the terms of the familiar mentalistic vocabulary—"belief", "desire", etc.—to express the corresponding descriptive theoretical concepts.

Return to Churchland's argument from reference. I suggested above that Churchland should not identify the common-sense mental-state concepts with the theoretical concepts of folk psychology. Rather, he might *explicate* the common-sense mental-state concepts with the descriptive theoretical concepts. Importantly, it follows that the implicit assumption—that the relevant theory of reference for the descriptive theoretical concepts is correct—is justified: by Referential Control, Churchland is permitted to *stipulate* that the extensions of those concepts are fixed descriptively by the laws of folk psychology. As such, if it really is a serious possibility that folk psychology is false, it follows that it is a serious possibility that the extensions of those concepts are *empty*. Moreover, by Explicandum Replacement (and given the fact that the argument concerns human psychology), Churchland ought to use the terms of the familiar mentalistic vocabulary to express the descriptive theoretical concepts. So, if he uses the familiar mentalistic vocabulary at all, he ought to express the conclusion of the argument along the following lines:

It is a serious possibility that there are no beliefs, desires, etc.

The conclusion of Churchland's argument, then, is retained.

One might object to this version of Churchland's argument from reference. In particular, one might object that Churchland would not be permitted to conclude that there are no beliefs, desires, etc., on the basis of just *one* explication: there might be an alternative explication of the common-sense mental-state concepts whose explicata are *non-empty*. Given such an explication, Churchland should presumably draw the conclusion that there *are* beliefs, desires, etc.

There are a couple of points to make about this. First, it does not suffice as an objection simply to note that there *might* be an alternative explication of the

common-sense mental-state concepts. The objector would have to argue that there are alternative, non-empty explicata for the common-sense mental-state concepts, which better satisfy (I)–(IV) than the descriptive theoretical concepts.

Second, Churchland is unlikely to accept that there are such alternative explicata. Churchland (1981: 72–76) offers a number of considerations in favour of the conclusion that nothing comes close to satisfying the laws of folk psychology. And, if that conclusion is right, then any non-empty explicata will either fail to satisfy criterion (I), which demands similarity to explicanda, or criterion (III), which demands fruitfulness. Suppose, first, that the non-empty explicata feature in generalisations that underpin explanations of human behaviour. Then, for Churchland, those explanations must be quite unlike the everyday explanations in which the common-sense mental-state concepts feature—otherwise we would expect the common-sense mental-state concepts to likewise be non-empty. As such, for Churchland, the non-empty explicata would in this key sense be dissimilar to our common-sense mental-state concepts and would thus fail to satisfy (I). But if the non-empty explicata did not feature in any generalisations that underpin explanations of human behaviour, then they would not be fruitful in Carnap's sense; they would not satisfy criterion (III). Either way, from Churchland's perspective, it is unlikely that there are non-empty explicata for the common-sense mental-state concepts that better satisfy (I)–(IV) than the descriptive theoretical concepts.

So Churchland can argue that, given the explication defence of arguments from reference, his conclusion (that it is a serious possibility that there are no beliefs, desires, etc.) stands.

4.4 Loose Ends

That is the explication defence of arguments from reference. Recall that, to respond to Machery et al.'s critique, the proponent of an argument from reference must provide a justification for her assumption that a particular theory of reference is correct. I have argued that, in principle and in at least two principal cases, this justification can be provided by appeal to explication. This is possible due to two key features of explication. First, given Referential Control, one can stipulate a theory of reference for an explicatum. Second, given Explicandum Replacement, one ought for theoretical purposes to use the explicatum in place of the explicandum, by using the relevant familiar term to express the explicatum rather than the explicandum. Post-explication, these features justify the assumption that the relevant familiar term is subject to the stipulated theory of reference. And, thus, the proponent of an argument from reference can use explication to *justify* her assumption that a particular theory of reference is correct.

There are some loose ends to tie up. First, I have focused on two examples that involved the assumption of a *descriptivist* theory of reference. This similarity is incidental: the examples were chosen because, first, I take Andreasen's argument to be particularly well suited to the explication defence and, second, the eliminativists' argument from reference is the example that Machery et al. spell out in most depth (Mallon et al. 2009: 334–336). For completeness, let me briefly consider another of

their examples, which involves the assumption of a *causal–historical* theory of reference.

Boyd (1988) argues that "developments in realist philosophy of science, together with related 'naturalistic' developments in epistemology and philosophy of language, can be employed in the articulation and defense of moral realism" (p. 308). For example, Boyd argues that, if a causal-historical theory of reference can "be extended to the analysis of moral language" (p. 325), then, in close analogy to scientific knowledge, one can give a plausible account of moral knowledge as a posteriori. Boyd goes on to sketch an account of the moral good as a homeostatic cluster of properties that satisfy important human needs (pp. 329–331), and explains why we might think that the cluster stands as the relevant causal-historical source of uses of "good" (pp. 336–338). As Mallon et al. write, "a significant epistemological position in ethics is [thus] derived from a specific theory about the reference of […] moral terms" (2009: 337).

To apply the explication defence, one might amend Boyd's position along the following lines: rather than assuming that the causal-historical theory of reference can be used to *analyse* moral language, one might offer an *explication* of moral language such that the explicata are subject to the causal-historical theory of reference. For example, one might suggest an explicatum for the common-sense concept *good* such that the explicatum denotes whatever stands as the relevant causal-historical source of ordinary uses of the common-sense concept. Boyd's position and arguments are naturally reworked to support the claim that such an explicatum is: (I) similar in relevant respects to the explicandum, as it is "plausible that the homeostatic cluster of fundamental human goods has, to a significant extent, regulated the [actual] use of the term 'good'" (p. 336); (II) precise, at least insofar as the invocation of the causal-historical theory of reference serves to fix an extension for the explicatum; and (III) fruitful, as the invocation of the causal-historical theory will ensure that the explicatum denotes a property that is *natural* in the same sense as "healthy" and "species" denote properties that are natural (pp. 322–325, 329–331). (I put aside the question of simplicity here.)

Given such an explication, Explicandum Replacement would imply that the explicatum should be used in place of the explicandum in relevant theoretical contexts, presumably including those in which one is doing metaethics. This justifies the assumption of a causal-historical theory of reference for "good", allowing one to subsequently draw Boyd's conclusion that there is a plausible account of moral knowledge as a posteriori. Thus, the explication defence of arguments from reference can be applied to arguments from reference that involve the assumption of a *causal-historical* theory of reference.

The second loose end: one might wonder whether my appeal to the method of explication is really required to respond to Machery et al.'s critique. I have appealed to that method in part to justify the stipulation of a theory of reference for "race", "belief", "desire", and so on. However, one might be tempted to suggest that a theory of reference could be stipulated for such terms without any such appeal.[21] For example: perhaps it would have been viable for Churchland to simply stipulate

[21] An anonymous reviewer for this journal makes a suggestion along these lines.

that he would use "belief", "desire", etc., as subject to a descriptivist theory of reference (with the laws of folk psychology acting as reference-fixing descriptions), and then argue that, in that sense, it is a serious possibility that there are no beliefs, desires, etc. Certainly, there seems to be no problem in performing such a stipulation; Churchland could have declared "Let 'belief', 'desire', etc., denote whatever (if anything) comes sufficiently close to satisfying the laws of folk psychology", and gone on from there.

The underlying difficulty facing this suggestion is that stipulations *per se* have no *normative* force: they do not vindicate any analogue of Explicandum Replacement. Thus, while I might be able to stipulate that I will use "belief" in any number of ways, it does not follow that I or anybody else *ought* to use "belief" in that way. For example, suppose that Churchland had stipulated a descriptivist theory of reference for "belief", "desire", etc., as suggested in the above paragraph, before expressing his conclusion by writing "it is a serious possibility that there are no beliefs, desires, etc." Then it would be reasonable to charge Churchland with obfuscating the issue; one might criticise Churchland for arguing merely that it is a serious possibility that nothing comes close to satisfying the laws of folk psychology, but misleadingly representing that conclusion, in an *ad hoc* fashion, as if it concerned beliefs, desires, etc. Such a charge is inappropriate, however, if the theory of reference is stipulated as part of an explication, as I have suggested. Explications seek to *refine* our concepts so that we *ought* to use appropriate explicata in place of explicanda when discussing relevant subject matter—as captured by Explicandum Replacement. Thus, when one deploys the explication defence of arguments from reference, one seeks to provide explicata that ought to be used in place of the explicanda; this justifies one's use of terminology, and gives other theorists reason to use it accordingly.

The third loose end: the kinds of objection that I raised in Sect. 3 do not gain traction with respect to the explication defence. First, as explications are stipulative, they are not subject to confirmation by appeal to experimental data concerning (folk or expert) semantic intuitions. Second, explications provide a positive justification for the assumption of a particular theory of reference. And, third, the factors that justify the assumption of a theory of reference are plausibly relevant to the conclusion of the argument from reference. For example, the question of which concepts are appropriate to deploy when theorising about human biology is plausibly very relevant to the substantive biological issue of whether race is biologically real. Likewise, the question of which concepts are appropriate to deploy when theorising about human psychology is plausibly very relevant to the substantive psychological issue of whether there are beliefs, desires, etc. Substantive issues about a given subject matter are not independent of how we should explicate the common-sense concepts that pertain to that subject matter.

Finally, I note that the explication defence may not work for *every* argument from reference. Consider an argument from reference that assumes a theory of reference, R, for a term that is typically used to express the common-sense concept c. The explication defence can be used just in case there is a good explication of c, such that it is stipulated that R applies to the explicatum. For only then is one justified in assuming that R is true of c. Whether there is a good explication will

depend on the specifics of the case, so I will not say anything general about the issue here. Regardless, as things stand, the explication defence can be used to respond to Machery et al.'s critique for at least two of their principal targets (and perhaps a third)—and I see no reason for thinking that it cannot be used in defence of other arguments from reference.

5 Concluding Remark

Machery et al.'s critique of arguments from reference is unsuccessful. Arguments from reference seek to establish substantive conclusions. And, as such, the concepts deployed in such arguments should be appropriate for theoretical inquiry. To this end, however, our common-sense concepts are often sub-optimal. To overcome the limitations of common-sense, theorists engineer *new* concepts that are better suited to the task. The semantics of these new concepts is not subject to empirical confirmation or disconfirmation through ordinary linguistic inquiry; we do not examine the intuitions of competent speakers to establish the extension of such concepts. Rather, theorists *stipulate* a theory of reference for the new concepts. And thus, when theorists deploy those concepts for theoretical purposes, it is legitimate for them to assume that the theory of reference is correct. Given a good explication of the relevant concepts, the assumption in an argument from reference that a theory of reference is correct is justified.

Acknowledgements For helpful comments and discussion I would like to thank James Andow, Emma Borg, Michael Devitt, Anthony Everett, Jumbly Grindrod, Nat Hansen, Krzysztof Posłajko, Kathy Puddifoot, two helpful referees for this journal, and audiences at PhiLang 2015, the Bucharest Colloquium in Analytic Philosophy 2015, and at the University of Reading. I gratefully acknowledge the financial support of the Analysis Trust, who have funded this research.

References

Andreasen, R. (1998). A new perspective on the race debate. *British Journal for the Philosophy of Science, 49*(2), 199–225.

Andreasen, R. (2000). Race: Biological reality or social construct? *Philosophy of Science, 67,* S653–S666.

Appiah, A. (1985). The uncompleted argument: Du Bois and the illusion of race. *Critical Inquiry, 12*(1), 21–37.

Boyd, R. (1988). How to be a moral realist. In G. Sayre-McCord (Ed.), *Essays on moral realism*. Ithaca: Cornell University Press.

Bradley, P., & Tye, M. (2001). Of colors, kestrels, caterpillars, and leaves. *Journal of Philosophy, 98*(9), 469–487.

Braisby, N., Franks, B., & Hampton, J. (1996). Essentialism, word use, and concepts. *Cognition, 59*(3), 247–274.

Carnap, R. (1950). *Logical foundations of probability*. Chicago: The University of Chicago Press.

Carnap, R. (1963). Replies and systematic expositions. In P. Schilpp (Ed.), *The philosophy of Rudolf Carnap* (pp. 859–1013). Lasalle, IL: Open Court.
Cavalli-Sforza, L. (1991). Genes, peoples, languages. *Scientific American, 265,* 104–110.
Churchland, P. M. (1979). *Scientific realism and the plasticity of mind*. Cambridge: Cambridge University Press.
Churchland, P. M. (1981). Eliminative materialism and the propositional attitudes. *Journal of Philosophy, 78*(2), 67–90.
Cohnitz, D., & Haukioja, J. (2015). Intuitions in philosophical semantics. *Erkenntnis, 80*(3), 617–641.
Deutsch, M. (2009). Experimental philosophy and the theory of reference. *Mind and Language, 24*(4), 445–466.
Devitt, M. (2006). *Ignorance of language*. Oxford: Oxford University Press.
Devitt, M. (2011). Experimental semantics. *Philosophy and Phenomenological Research, 82*(2), 418–435.
Devitt, M. (2012a). Whither experimental semantics? *Theoria, 27*(1), 5–36.
Devitt, M. (2012b). Semantic epistemology: Response to Machery. *Theoria, 27*(2), 229–233.
Genone, J., & Lombrozo, T. (2012). Concept possession, experimental semantics, and hybrid theories of reference. *Philosophical Psychology, 25*(5), 717–742.
Hansen, N. (2015). Experimental philosophy of language. *Oxford Handbooks Online*. doi:10.1093/oxfordhb/9780199935314.013.53.
Ichikawa, J., Maitra, I., & Weatherson, B. (2012). In defense of a Kripkean dogma. *Philosophy and Phenomenological Research, 85*(1), 56–68.
Justus, J. (2012). Carnap on concept determination: Methodology for philosophy of science. *European Journal for Philosophy of Science, 2,* 161–179.
Jylkkä, J., Railo, H., & Haukioja, J. (2009). Psychological essentialism and semantic externalism: Evidence for externalism in lay speakers' language use. *Philosophical Psychology, 22*(1), 37–60.
Kitcher, P. (2008). Carnap and the caterpillar. *Philosophical Topics, 36*(1), 111–127.
Kripke, S. (1980). *Naming and necessity*. Cambridge, MA: Harvard University Press.
Lam, B. (2010). Are Cantonese-speakers really descriptivists? Revisiting cross-cultural semantics. *Cognition, 115,* 320–329.
Ludlow, P. (2014). *Living words*. Oxford: Oxford University Press.
Ludwig, K. (2007). The epistemology of thought experiments: First-person versus third-person approaches. *Midwest Studies in Philosophy, 31,* 128–159.
Lycan, W. (1988). *Judgement and justification*. Cambridge: Cambridge University Press.
Machery, E. (2012a). Expertise and intuitions about reference. *Theoria, 27*(1), 37–54.
Machery, E. (2012b). Semantic epistemology: A brief response to Devitt. *Theoria, 27*(2), 223–227.
Machery, E., Deutsch, M., Mallon, R., Nichols, S., Sytsma, J., & Stich, S. (2010). Semantic intuitions: Reply to Lam. *Cognition, 117,* 361–366.
Machery, E., Deutsch, M., & Sytsma, J. (2015). Speaker's reference and cross-cultural semantics. In A. Bianchi (Ed.), *On reference*. Oxford: Oxford University Press.
Machery, E., Mallon, R., Nichols, S., & Stich, S. (2004). Semantics, cross-cultural style. *Cognition, 92,* B1–B12.
Machery, E., Mallon, R., Nichols, S., & Stich, S. (2013). If folk intuitions vary, then what? *Philosophy and Phenomenological Research, 86*(3), 618–635.
Machery, E., Olivola, C., & de Blanc, M. (2009). Linguistic and metalinguistic intuitions in the philosophy of language. *Analysis, 69*(4), 689–694.
Mackie, J. L. (1977). *Ethics: Inventing right and wrong*. Harmondsworth: Penguin.
Maher, P. (2007). Explication defended. *Studia Logica, 86,* 331–341.
Mallon, R. (2006). 'Race': Normative, not metaphysical or semantic. *Ethics, 116,* 525–551.
Mallon, R., Machery, E., Nichols, S., & Stich, S. (2009). Against arguments from reference. *Philosophy and Phenomenological Research, 79*(2), 332–356.
Martí, G. (2009). Against semantic multi-culturalism. *Analysis, 69*(1), 42–48.
Maund, B. (2011). Colour eliminativism. In L. Nolan (Ed.), *Primary and secondary qualities: The historical and ongoing debate* (pp. 362–385). Oxford: Oxford University Press.
Nichols, S., Ángel Pinillos, N., & Mallon, R. (2016). Ambiguous reference. *Mind, 125*(497), 145–175.
Schupbach, J. (2015). Experimental explication. *Philosophy and Phenomenological Research*. doi:10.1111/phpr.12207.
Shepherd, J., & Justus, J. (2015). X-phi and Carnapian explication. *Erkenntnis, 80*(2), 381–402.

Sytsma, J., & Livengood, J. (2011). A new perspective concerning experiments on semantic intuitions. *Australasian Journal of Philosophy, 89*(2), 315–332.

Sytsma, J., Livengood, J., Sato, R., & Oguchi, M. (2015). Reference in the land of the rising sun: A cross-cultural study on the reference of proper names. *Review of Philosophy and Psychology, 6*(2), 213–230.